Virtual Biology Corporation

Revolutionizing Drug Discovery Through AI-Powered Virtual Cell Models

Confidential White Paper

November 4, 2024

Executive Summary

Problem: Current virtual cell models are bottlenecked by single-cell RNA sequencing (scRNA-seq) data, hindering scalability and accuracy.

Solution: We are developing cross-species virtual cell models that leverage genomic sequences and comparative genomics to predict gene expression and cell types directly from DNA, bypassing the data bottleneck.

Opportunity: Our approach will revolutionize drug discovery and biological research by enabling accurate *in silico* simulations of genetic and pharmacological perturbations, reducing costs, and accelerating development timelines.

Risk: While facing significant competition in pharmaceuticals, our cross-species platform enables rapid expansion into the underserved veterinary and agricultural markets, reducing market-entry barriers and diversifying revenue streams.

1 Introduction

The development of virtual cell models holds immense potential for transforming drug discovery and biological research. However, existing models face two critical challenges:

- Data scarcity and expense of single-cell RNA sequencing
- Limited clinical trial populations failing to capture human genetic diversity

With clinical trials typically involving just thousands of participants meant to represent billions of people, and a resulting 90% failure rate costing over $2 billion per drug, there's an urgent need for a paradigm shift. Our platform addresses both challenges by predicting cellular responses directly from genomic sequences, enabling truly personalized drug development that accounts for individual genetic variations across the entire human population.

2 The Science

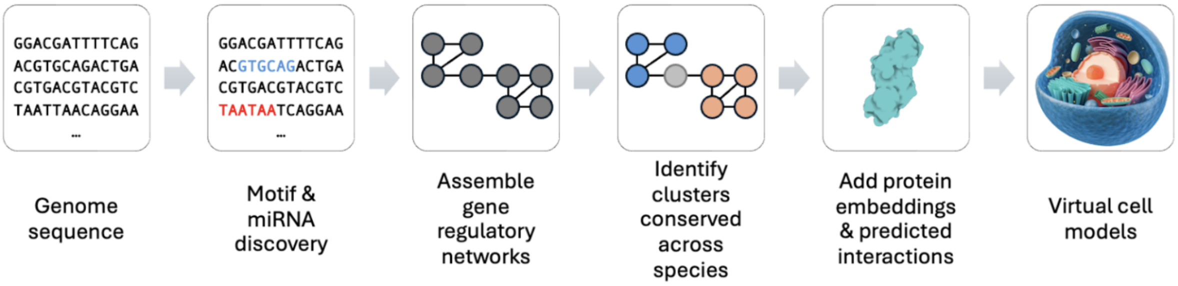

Figure 1: **Virtual Biology Process Overview.** Our platform leverages genomic sequences and comparative genomics to predict gene expression and cell types directly from DNA, creating a scalable approach to drug discovery.

Our breakthrough approach harnesses biology's digital foundation: DNA. Just as GPT models learned to understand human language, our genomic language models (gLMs) learn to read and interpret the genetic code that builds every living organism. By analyzing DNA sequences across species, we identify the conserved 'genetic programs' that control cell behavior—essentially decoding nature's own programming language.

While current bio AI remains in its infancy (comparable to GPT-1), we're pioneering the digitization of the gene-to-cell pipeline through two key innovations:

Genomic Language Models (gLMs):

Our models learn regulatory patterns directly from DNA sequences, predicting gene expression and cell behavior with unprecedented accuracy. This bypasses expensive RNA sequencing, creating a unique and cost-effective data moat.

Cross-Species Analysis:

By studying genetic programs across multiple species, we leverage billions of years of evolutionary optimization to build more robust and generalizable models. This enables:

- Generation of personalized predictions for any individual's genome
- Extension of our platform to any species of interest
- Construction of cell-type specific models without expensive experimental data

This digital biology platform represents a fundamental shift from traditional approaches, positioning us at the forefront of the bio AI revolution—where understanding one genome helps us understand them all.

3 The Method

We are building knowledge graphs that represent interacting genes and their relationships, capturing cause-effect dynamics within the cell. By integrating gene annotations, regulatory motifs, and protein structures into these graphs, we can simulate genetic and pharmacological perturbations by modulating specific interactions within the network.

Our strategy employs graph neural networks (GNNs) to classify these knowledge graphs through supervised learning, predicting cellular outcomes based on historical data. This approach bypasses the data bottleneck of scRNA-seq, enabling scalable and accurate virtual cell modeling across diverse species. Our classifiers can also be trained from hundreds of thousands of public genome sequences.

By comparing knowledge graphs across dozens of related species (e.g., vertebrates), we can enhance the predictive power of our models. This method has been inspired by the success of tools like SATURN, which demonstrate that incorporating more species leads to better models.